수집된 데이터 및 데이터 전처리 문서

목차

내용

1. 데이터 설명

- 수집 방법: 크롤링, 오픈 데이터 소스
- 수집 기간: 12/23 - 1.3
- 도구 및 기술: Selenium, BeautifulSoup
- 데이터 소스: 네이버지도, 공공데이터포탈 (서울관광재단), LOCALDATA

- 네이버 지도 크롤링 데이터

크롤링 키워드	음식점, 술집, 카페	동 단위 총 180개 키워드 검색 예시) 강남구 논현동 카페
총 데이터 수	23693개	
컬럼 수	17 개	store_name,category,new_open,rating,visited_review,directions_text,store_id,address,blog_review,phone_num,business_hours,info,convenience_facilities_and_services,Parking,sns,review_like_this_part_output,menu
형식	csv	

- 오픈 소스 데이터

데이터 이름	사이트 이름	특징	링크	파일크기	갱신 년월	데이터 수	컬럼
한국보건산업진흥원_외국인환자 유치기관 현황	공공데이터 포탈	한국보건산업진흥원 외국인환자 유치기관 현황 자료입니다.(유치기관 상태, 유치기관명, 유치기관유형, 지역, 유치대상 국가 등)	https://www.data.go.kr/data/3050000/fileData.do	1.5 MB	2024-07	1738개	상호,대표자,기관구분,주소,타겟국가
서울시 의료관광허가 의료기관 정보(한국어)	공공데이터 포탈(서울관광재단)	의료관광허가가된 의료기관의 상호 및 주소, 타겟국가 등 관련 정보 제공(한국어)	https://data.seoul.go.kr/dataList/OA-12973/S/1/datasetView.do	226 KB	2024-04		
서울시 관광음식(서울의 맛집 탐방. 한식, 중식, 일식, 아시아식, 서양식, 주점 등 메뉴별 맛집, 카페&디저트, 채식, 할랄 등 특화 식당까지 정보 제공)	공공데이터 포탈(서울관광재단)		https://data.seoul.go.kr/dataList/OA-21054/S/1/datasetView.do	453 KB	매월 4일	725개	상호명,콘텐츠URL,주소,신주소,전화번호,웹사이트,운영시간,교통정보,홈페이지 언어,대표메뉴

서울시 관광명소 (서울의 랜드마크, 고궁, 역사적 장소, 오래가게, 미술관, 박물관 등 놓칠 수 없는 서울의 명소 장소 정보)	공공데이터포탈 (서울관광재단)		https://data.seoul.go.kr/dataList/OA-21050/S/1/datasetView.do	219 KB	매월 4일	280개	상호명,콘텐츠URL,주소,신주소,전화번호,팩스번호,웹사이트,운영시간,운영요일,휴무일,교통정보,태그,장애인편의시설
서울특별시_관광 쇼핑 (서울의 쇼핑몰, 백화점, 면세점, 마켓, 뷰티샵, 한류 및 관광상품 등을 세계 각종 언어로 상호명, 주소 위치정보 등 공개합니다.)	공공데이터포탈 (서울관광재단)		https://data.seoul.go.kr/dataList/OA-21053/S/1/datasetView.do	94 KB	매월 4일	153개	상호명,콘텐츠URL,주소,신주소,전화번호,웹사이트,운영시간,교통정보
서울특별시_관광 문화 (주요 언어로 관광시설,문화시설, 휴식시설, 스포츠시설, 놀이공원, 테마파크, 체험공간 등 익사이팅한 서울의 놀거리들 정보를 안내)	공공데이터포탈 (서울관광재단)		https://data.seoul.go.kr/dataList/OA-21052/S/1/datasetView.do	116 KB	매월 4일	13개	상호명,콘텐츠URL,주소,신주소,전화번호,웹사이트,운영시간,운영요일,휴무일,교통정보,홈페이지 운영 언어,유모차 대여 여부

서울특별시_관광 자연(도심 속에서 자연의 힐링을 느낄 수 있는 서울의 산, 강, 계곡, 섬, 공원, 정원, 식물원 등 자연 관광명소들을 언어별로 소개합니다.)	공공데이터 포탈 (서울관광재단)		https://data.seoul.go.kr/dataList/OA-21051/S/1/datasetView.do	56 KB	매월 4일	33개	상호명,콘텐츠URL,주소,신주소,전화번호,웹사이트,운영시간,운영요일,휴무일,교통정보
서울특별시_관광거리 정보 (서울특별시의 주요 관광거리에 대한 공식 명칭 및 주소 등 정보를 제공합니다.(한국어) 검색키워드,최종표기명,지번주소,법정시,법정구,법정동,위치정보 등을 제공합니다.)	공공데이터 포탈 (서울관광재단)	경도 위도 포함	https://data.seoul.go.kr/dataList/OA-12929/S/1/datasetView.do	25 KB	2016-02	54개	최종 표기명,지번 주소,중심 좌표 X,중심 좌표 Y
서울특별시_ 관광안내소 (표준 데이터) 서울특별시 관광안내소 정보(안내소소개, 부가서비스 정보,	공공데이터 포탈 (서울관광재단)	경도 위도 포함	https://data.seoul.go.kr/dataList/OA-20350/S/1/datasetView.do	7 KB	2024-12	8개	관광안내소명,안내소위치명,안내소소개,부가서비스정보,휴무일,운영시작시각(하절기),

휴무일, 운영시간, 근무인원수, 외국어안내 가능여부, 전화번호 등)를 제공합니다							운영종료시각(하절기),운영시작시각(동절기),운영종료시각(동절기),안내가능외국어,안내소전화번호,소재지도로명주소,소재지지번주소,홈페이지주소,위도,경도
관광숙박업	LOCALDATA	경도 위도 데이터 포함	https://www.localdata.go.kr/data/dataView.do	186 KB		474개	인허가일자,소재지전화,소재지전체주소,도로명전체주소,도로명우편번호,사업장명,좌표정보X(EPSG5174),좌표정보Y(EPSG5174),관광숙박업상세명

3. 전처리 순서

- 네이버
  - 데이터 병합
  - 중복 제거
  - 결측치 처리
  - 데이터 변환
  - 벡터 스토어 임베딩
- 오픈 소스 데이터

- 행정구 단위 데이터 분리
- 데이터 변환
- 행정구 별 데이터 병합
- 벡터 스토어 임베딩

4. 전처리 단계별 세부 내용

- 네이버 지도
  - 데이터 병합 **:** 수집한 데이터를 한번에 처리하기 위해 각 종류별 데이터 병합
    - 키워드별(행정동 단위 키워드별) 총 180개의 데이터 병합
  - 중복 제거
    - 수집된 csv 파일을 pandas 라이브러리의 dataframe 으로 변경
    - store_id 컬럼을 기준으로 drop_duplicates() 함수를 이용하여 중복 제거
    - 오픈소스 데이터는 중복 데이터 없음
      결과 : 네이버 지도 크롤링 데이터 47700개 -> 23693개
  - 결측치 처리
    - 네이버지도 데이터 23693개 중 store_name 컬럼 816개 결측
    - 3.4% 결측치 존재
    - pandas 라이브러리의 dropna() 함수를 통해 결측치 제거
      결과 : 네이버 지도 크롤링 데이터 23693개 -> 22877개
  - 데이터 변환
    - csv 파일에서 json 형식으로 변환
      - `to_json.()`함수를 사용하여 변환
  - 벡터 스토어 임베딩
    - faiss 라이브러리를 사용해서 json 형식의 데이터를 벡터 스토어로 임베딩
    - 벡터 스토어 구조

meta data	page content
store_name, store_id	store_name, category, new_open, rating, visited_review, directions_text, store_id, address, blog_review, phone_num, business_hours, info, convenience_facilities_and_services, Parking, sns, review_like_this_part_output,menu

- 오픈 소스 데이터

- 행정구 단위 데이터 분리
  - 9개의 데이터를 각각 강남구, 종로구, 중구, 용산구 지역별로 데이터 필터링 하여 파일 분리,
    파일 수 : 9개 → 36개
    - 9개의 데이터 : 음식점, 관광거리, 명소, 외국인 유치 병원, 관광 안내소, 문화 시설, 자연 체험, 숙박업, 쇼핑거리
- 데이터 변환
  - csv 파일에서 json 형식으로 변환
    - `to_json.()`함수를 사용하여 변환
- 행정구 별 데이터 병합
  - 36개의 데이터 강남구, 종로구, 중구, 용산구 지역별로 파일 병합,
    파일 수 : 36개 → 4개
- 벡터 스토어 임베딩
  - faiss 라이브러리를 사용해서 json 형식의 데이터를 벡터 스토어로 임베딩
  - 벡터 데이터 구조

- 음식점

meta data	page content
상호명	상호명, 콘텐츠URL, 주소, 신주소, 전화번호, 웹사이트, 운영시간, 교통정보, 홈페이지 언어, 대표메뉴

- 관광 거리

meta data	page content
최종 표기명	최종 표기명, 지번 주소, 중심 좌표 X, 중심 좌표 Y

- 명소

meta data	page content
상호명	상호명, 콘텐츠URL, 주소, 신주소, 전화번호, 팩스번호, 웹사이트, 운영시간, 운영요일, 휴무일, 교통정보, 태그, 장애인편의시설

- 관광 안내소

meta data	page content
관광안내소명	관광안내소명, 안내소위치명, 안내소소개, 부가서비스정보, 휴무일 ,운영시작시각(하절기) , 운영종료시각(하절기), 운영시작시각(동절기), 운영종료시각(동절기), 안내가능외국어, 안내소전화번호 ,소재지도로명주소, 소재지지번주소, 페이지주소, 위도, 경도

- 문화 시설

meta data	page content
상호명	상호명, 콘텐츠URL, 주소, 신주소, 전화번호, 웹사이트, 운영시간, 운영요일, 휴무일, 교통정보, 홈페이지 운영 언어, 유모차 대여 여부

- 외국인 유치 병원

meta data	page content
상호	상호, 대표자, 기관구분, 주소, 타겟국가

- 쇼핑센터(거리)

meta data	page content
상호명	상호명, 콘텐츠URL, 주소, 신주소, 전화번호, 웹사이트, 운영시간, 교통정보

- 자연체험

meta data	page content
상호명	상호명, 콘텐츠URL, 주소, 신주소, 전화번호, 웹사이트, 운영시간, 운영요일, 휴무일 , 교통정보

- 호텔

meta data	page content
사업장명	인허가일자, 소재지전화, 소재지전체주소, 도로명전체주소 , 도로명우편번호, 사업장명, 좌표정보X(EPSG5174), 좌표정보Y(EPSG5174), 관광숙박업상세명

5. 결과

- 데이터 크기 :
    - 네이버 크롤링 데이터
        - json : 63.7 MB
        - faiss 벡터 스토어 : 252.6MB
    - 오픈 소스 데이터 -
        - json : 1.8MB
        - faiss 벡터 스토어 : 23.5MB
- 데이터 수:
    - 네이버 크롤링 데이터 : 22877개
    - 오픈 소스 데이터 : 3478개
- 종합
    - 총 데이터 크기
        - json : 65.5MB
        - faiss 벡터 스토어 : 276.1MB
    - 총 데이터 수
        - 26355개